

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-015832
 (43)Date of publication of application : 22.01.1999

(51)Int.CI. G06F 17/30

(21)Application number : 09-168267
 (22)Date of filing : 25.06.1997

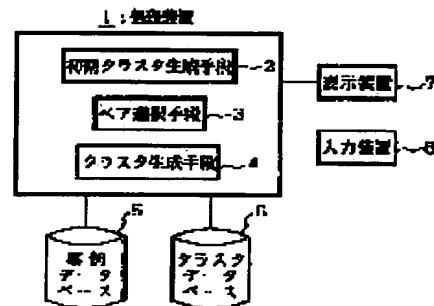
(71)Applicant : FUJITSU LTD
 (72)Inventor : OOTA TADAKO
 YUGAMI NOBUHIRO
 OKAMOTO AOSHI
 SATO OSAMU

(54) CLUSTER GENERATION DEVICE AND RECORDING MEDIUM

(57)Abstract:

PROBLEM TO BE SOLVED: To generate a cluster hierarchy permitting that one example belongs to plural clusters by means of shortening cluster generation time by considering the cluster whose number of appearing examples in common is large in a cluster set to be highly associated and generating the new cluster from the plural clusters.

SOLUTION: An initial cluster generation means 2 generates the clusters for respective attribute values based on the given example, which is read from an example data base 5. A pair selection means 3 selects the pair of the clusters in the generated clusters. A cluster generation means 4 counts the number of the examples contained in common on the selected pair, generates the new cluster from the pertinent pair of plural clusters in the largest case and stores it in a cluster data base 6. The cluster whose number of the appearing examples in common is large in the cluster set is considered to be highly associated. The generation of the new cluster from the plural clusters is repeated and the cluster hierarchy is generated.



LEGAL STATUS

- [Date of request for examination]
- [Date of sending the examiner's decision of rejection]
- [Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
- [Date of final disposal for application]
- [Patent number]
- [Date of registration]
- [Number of appeal against examiner's decision of rejection]
- [Date of requesting appeal against examiner's decision of rejection]
- [Date of extinction of right]

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平11-15832

(43)公開日 平成11年(1999)1月22日

(51)Int.Cl.⁶

G 0 6 F 17/30

識別記号

F I

G 0 6 F 15/401

3 1 0 D

審査請求 未請求 請求項の数4 OL (全9頁)

(21)出願番号 特願平9-168267

(22)出願日 平成9年(1997)6月25日

(71)出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中4丁目1番
1号

(72)発明者 太田 唯子

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(72)発明者 潟上 伸弘

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(74)代理人 弁理士 岡田 守弘

最終頁に続く

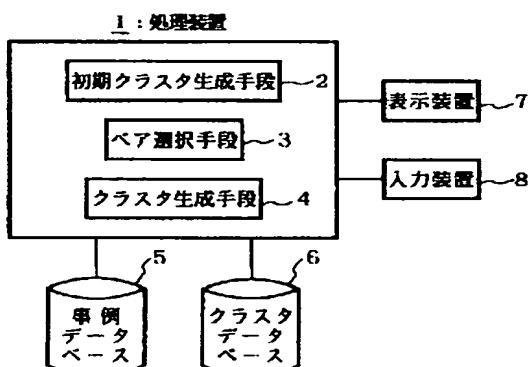
(54)【発明の名称】 クラスタ生成装置および記録媒体

(57)【要約】

【課題】 本発明は、与えられた事例をもとにクラスタを生成するクラスタ生成装置に関し、各属性の各値についてその値を持つ事例を集めてクラスタを生成し、クラスタ集合中で共通して出現する事例の数の多いクラスタは関連が高いと見做して複数クラスタから新たなクラスタを生成することを繰り返してクラスタ階層を生成し、クラスタ生成時間を短くして1つの事例が複数のクラスタに属することを許すクラスタ階層を生成することとする。

【解決手段】 与えられた事例をもとに属性値毎にクラスタを生成する手段と、生成されたクラスタ中の複数のクラスタに共通して含まれる事例の数が多い場合に当該複数のクラスタから新たなクラスタを生成する手段とを備えるように構成する。

本発明のシステム構成図



1

【特許請求の範囲】

【請求項1】与えられた事例をもとにクラスタを生成するクラスタ生成装置において、

与えられた事例をもとに属性値毎にクラスタを生成する手段と、

上記生成されたクラスタ中の複数のクラスタに共通して含まれる事例の数が多い場合に当該複数のクラスタから新たなクラスタを生成する手段とを備えたことを特徴とするクラスタ生成装置。

【請求項2】上記複数のクラスタから新たなクラスタを生成するとして、複数のクラスタの和集合を新たなクラスタとすることを特徴とする請求項1記載のクラスタ生成装置。

【請求項3】上記複数のクラスタから新たなクラスタを生成するとして、複数のクラスタの積集合を新たなクラスタとすることを特徴とする請求項1記載のクラスタ生成装置。

【請求項4】与えられた事例をもとに属性値毎にクラスタを生成する手段と、上記生成されたクラスタ中の複数のクラスタに共通して含まれる事例の数が多い場合に当該複数のクラスタから新たなクラスタを生成する手段として機能するプログラムを格納した記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、与えられた事例をもとにクラスタを生成するクラスタ生成装置に関するものである。

【0002】事例を分類するクラスが与えられていない場合、どのようなクラスに分類するかが重要な問題となる。クラスタリングは、そのような場合、クラスそのもの、または部分集合となるようなクラスを生成する技術である。一般に、クラスタは似た事例の集合である。例えばアンケートデータなどから、似た人達のグループを検出するのに利用できる。このクラスタの生成においては、クラスタ間の類似性の計り方、クラスタの生成の仕方によって、得られるクラスタ階層の性能が変わってくる。本発明は、そのクラスタ間の類似性の計り方、生成の仕方に関するものである。

【0003】

【従来の技術】従来、クラスタ生成の手法には、逐次的手法と一括的手法がある。逐次的手法は、クラスタリングされる事例を、一度ではなく、少しずつ時間をおって与えられる。新しい事例が与えられた時、それまでに作られたクラスタを始めから作り直すのではなく、それに変更を加える操作を行う。そのため、クラスタ生成に必要とされる時間は短いが、生成されたクラスタ階層の性能は、事例の入力順序に大きく依存する。

【0004】一括的手法は、事例が一度に与えられるため、入力順序の問題はないが、クラスタ生成に時間がかかる。例えば、代表的な手法としてWard法がある

2

が、この計算量は、全事例数の3乗である。

【0005】

【発明が解決しようとする課題】上述したようにクラスタ生成を従来の逐次的手法で行うと、新しい事例が与えられたときにクラスタ生成を迅速に行うことができるが、生成されるクラスタ階層が事例の入力順序に大きく依存してしまう問題がある。

【0006】また、従来の一括的手法で行うと、入力順序の問題はないが、クラスタ生成に時間がかかるという

10 問題がある。また、逐次的／一括的手法に共通して多くの手法は、1つの事例が1つのクラスにしか属さないようなクラスタ階層が生成される。ある問題を解決するために、複数の視点が要求される場合も多く、従来の1つの事例が1つのクラスにしか含まれないような階層を生成する方法では、各視点ごとに階層を生成する必要も生じてしまう。そこで、事例の入力順序によらず、階層生成に必要な時間が短く、更に、1つの事例が複数のクラスタに属することを許すようなクラスタ階層生成手法が望まれる。

20 【0007】本発明は、これらの問題を解決するため、各属性の各値についてその値を持つ事例を集めてクラスタを生成し、クラスタ集合中で共通して出現する事例の数の多いクラスタは関連が高いと見做して複数クラスタから新たなクラスタを生成することを繰り返してクラスタ階層を生成し、クラスタ生成時間を短くして1つの事例が複数のクラスタに属することを許すクラスタ階層を生成することを目的としている。

【0008】

【課題を解決するための手段】図1を参照して課題を解30 決するための手段を説明する。図1において、処理装置1は、図示外の記録媒体からプログラムを主記憶にロードイングして起動し各種処理を行うものであって、ここでは、初期クラスタ生成手段2、ペア選択手段3、およびクラスタ生成手段4などから構成されるものである。

【0009】初期クラスタ生成手段2は、クラスタを生成するものである。ペア選択手段3は、クラスタのペアを選択するものである。クラスタ生成手段4は、選択したペアについて共通出現事例数を計数し最も多い数のペアを新たなクラスタとして生成などするものである。

40 【0010】事例データベース5は、事例を格納したものである。クラスタデータベース6は、生成したクラスタを格納するものである。表示装置7は、クラスタを表示などするものである。

【0011】入力装置8は、各種指示やデータを入力するものである。次に、動作を説明する。初期クラスタ生成手段2が事例データベース5から読み出した与えられた事例をもとに属性値毎にクラスタを生成し、ペア選択手段3が生成されたクラスタ中の複数（例えば2つ）のクラスタのペアを選択し、クラスタ生成手段4が選択されたペアについて共通して含まれる事例の数を計数し最

50 大きな値をもつペアを新たなクラスタとして生成する。

3

も多い場合に当該ペアの複数のクラスタから新たなクラスタを生成し、クラスタデータベース6に格納するようしている。

【0012】この際、複数のクラスタの和集合を新たなクラスタを生成するようにしている。また、複数のクラスタの積集合を新たなクラスタを生成するようにしている。

【0013】従って、各属性の各値についてその値を持つ事例を集めてクラスタを生成し、クラスタ集合中で共通して出現する事例の数の多いクラスタは関連が高いと見做して複数クラスタから新たなクラスタを生成することを繰り返してクラスタ階層を生成することにより、クラスタ生成時間を短くして1つの事例が複数のクラスタに属することを許すクラスタ階層を生成することが可能となる。

【0014】

【発明の実施の形態】次に、図2から図5を用いて本発明の実施の形態および動作を順次詳細に説明する。

【0015】図2は、本発明の動作説明フローチャートを示す。これは、図1の構成の詳細な動作を説明するフローチャートである。図2において、S1は、事例集合を入力する。

【0016】S2は、パラメタKとクラスタ制限数Lを決定する。これは、右側に記載したように、パラメタKとクラスタ制御数Lを入力し、これをもとに右側に記載したように、 $K \times$ 標準偏差=標準偏差の正の実数倍の決定、およびクラスあるいはその部分集合(クラスタ)の総計の最大数をLとして決定する。

【0017】S3は、初期クラスタ階層を生成する。これは、S1で入力された事例集合について、後述する図3に示すように初期クラスタ階層を生成、即ち、後述する図3に示すように、入力された全事例を含むクラスタをCrootとし、葉クラスタはそれぞれ属性a_mの値がv

$$e_{ij} = (|C_i| \times |C_j|) / N$$

(1)

ここで、 $|C_i|$ 、 $|C_j|$ はクラスタi、jに含まれるそれぞれの事例数を表し、Nは全事例数を表す。このと

$$d_{ij} = \{ |C_i| \times |C_j| \times (N - |C_i|) \times (N - |C_j|) \}^{1/2} / \{ N^2 \times (N - 1) \}^{1/2}$$

(2)

となる。ここで、パラメタKを含む、以下のような式

(3) を満たす時に、その2つのクラスタの関連が高い

$$|C_i \cap C_j| > e_{ij} + K \cdot d_{ij}$$

(3)

ここで、左側は後述する図4の(e)の積集合を表す。

次に、下記の式(4)を計算する(S9)。

$$|C_i \cap C_j| / (|C_i| \times |C_j|)$$

(4)

ここで、左側の式は後述する図4の(e)の積集合を表し、右側の式は後述する図4の(d)の和集合を表す。

【0026】これら計算した結果から、式(3)を満たし、式(4)を最大化するクラスタペアを選択することにより、事例数が多いクラスタ同士が選択される傾向を緩和できる(S9、S10)。

4

m_v である事例を集めたクラスタとからなる初期クラスタ階層を生成する。

【0018】S4は、終了か判別する。これは、S5以下の処理について、クラスタ数がL個となるか、ペアとして選択するクラスタがなくなるまで繰り返すことが終了したか判別する。YESの場合には、終了する。NOの場合には、S5に進む。

【0019】S5は、クラスタペアを選択する。S6は、共通出現事例数を数える。これは、選択したペアの10クラスタについて、共通して出現する事例の数を数える。

【0020】S7は、推定共通事例数を計算する(後述する)。S8は、S6で数えた共通出現事例数がS7で計算した推定共通事例数よりも十分に大きいか判別する。YESの場合には、S9に進む。NOの場合には、S4に戻り繰り返す。

【0021】S9は、事例数の割合を計算する。S10は、一番大きい割合を記憶する。S11は、全ペア試行したか判別する。YESの場合には、S12で一番割合20の大きいペアを選択し、S13で新たなクラスタを生成し、S4に戻り繰り返す。一方、S11のNOの場合には、S4に戻り繰り返す。

【0022】ここで、S3ないしS13について以下に詳細に説明する。後述する図3に示すように、全ての事例を含むクラスタをCrootとして生成し、葉クラスタとして属性a_mの値がv_mである事例の集合を生成する(S3)。そして、共通して出現する事例数のクラスタのペアの関連度を評価する。その数が推定共通事例数よりも十分に大きい場合(S8のYESの場合)、関連が高いと見なす。推定共通事例数は、以下の式によって計算する。

【0023】

きの標準偏差は、

と見做され、選択候補となる(S8)。

【0024】

【0025】

【0027】そして、選択されたクラスタペアの和集合を新たなクラスタとして生成し、選択ペアの上位にリンクすることにより、後述する図4の(b)となる。また、選択されたクラスタペアの積集合を新たなクラスタとして生成し、選択ペアの下位にリンクすることにより、後述する図4の(c)となる。

【0028】以上のように、各属性値についてクラスタを生成し、関連の大きいクラスタペアから新たなクラスタを生成することを繰り返しクラスタ階層を生成することにより、クラスタ生成時間を短くして1つの事例が複数のクラスに属することを許すクラスタ階層を生成することが可能となる。以下順次詳細に説明する。

【0029】図3は、本発明の説明図(その1)を示す。これは、初期クラスタの概念を説明する図である。図3の(a)は、初期クラスタの概念図を示す。ここで、全ての事例を含むクラスタをCrootとして生成する。次に、葉クラスタとして属性amの値がvmvである事例の集合を図示のように生成する。

【0030】図3の(b)は、初期クラスタの具体例を示す。ここで、全ての事例を含むクラスタCrootとしてクラスタ“家具”を生成する。次に、葉クラスタとして、属性“脚の数”的値が“0”、“3”を持つものをクラスタ“脚の数=0”、“脚の数=3”として図示のように生成する。同様に、葉クラスタとして、属性“材質”的値が“木”を持つものをクラスタ“材質=木”として図示のように生成する。

【0031】以上のように全ての事例を含むクラスタCrootを生成し、葉クラスタとして属性の値毎に事例の集合を生成することによって、事例集合から初期クラスタを生成することが可能となる。

【0032】図4は、本発明の説明図(その2)を示す。これは、初期クラスタから関連の強い複数のクラスタから1つの新たなクラスタを生成するときの概念を説明する図である。

【0033】図4の(a)は、クラスタ階層中から選択されたペアの2つのクラスタCi、Cjを示す。図4の(b)は、和集合を新クラスタとする方法の例を示す。この2つのクラスタCi、Cjの和集合を新たなクラスタCmとする場合には、当該2つのクラスタCi、Cjの和集合の新たなクラスタCmを図示のようにペアの上位にリンク付けする。

【0034】図4の(c)は、積集合を新クラスタとする方法の例を示す。この2つのクラスタCi、Cjの積集合を新たなクラスタCmとする場合には、当該2つのクラスタCi、Cjの積集合の新たなクラスタCmを図示のようにペアの下位にリンク付けする。

【0035】図4の(d)は、和集合を模式的に示した図である。和集合(図4の(b)の場合)は、

$$|C_i| + |C_j|$$

で表現され、クラスタCiとクラスタCjとのそれぞれの斜線で示す部分の和となる。

【0036】図4の(e)は、積集合を模式的に示した図である。積集合(図4の(c)の場合)は、

$$|C_i \cap C_j|$$

で表現され、クラスタCiとクラスタCjとのそれぞれの斜線で示す重なる部分となる。

【0037】図5は、本発明のクラスタ階層例を示す。これは、図3の(b)の初期クラスタについて、既述した和集合の場合の新たなクラスタ、および既述した積集合の場合の新たなクラスタを生成した後のクラスタ階層例である。

【0038】例えば左側の和集合のクラスタは、クラスタ“脚の数=3”とクラスタ“材質=木”的2つのペアの和集合として新たなクラスタ“脚の数=3 and 材質=木”を生成して上位にリンクしたものである。

【0039】同様に、例えば右側の積集合のクラスタは、クラスタ“材質=プラスティック”とクラスタ“形=四角”的2つのペアの積集合として新たなクラスタ“材質=プラスティック and 形=四角”を生成して下位にリンクしたものである。

【0040】以上のように、クラスタ集合中から関連するペアを見つけ、その和集合／積集合を上位／下位にリンク付けすることにより、関連の大きいクラスタペアから新たなクラスタを生成することが可能となる。

【0041】図6および図7は、本発明のシステム動作説明フローチャートを示す。これは、既述した図2の詳細なシステム動作を説明するフローチャートであって、図2のS1ないしS13に対応する部分を左端に記載する。

【0042】図6において、S21は、事例集合を入力する(S1)。S22は、パラメタkと、クラスタ制限数lを決定する。S23は、m=0と初期設定する。これは、新たなクラスタCmを生成するときの変数mの値を初期化する。

【0043】S24は、i=0と初期設定する。S25は、j=0と初期設定する。S26は、属性aiが値viである全事例からなるクラスタCmを作る。

【0044】S27は、m=m+1する。S28は、j=(aiの取る値の数-1)か判別する。YESの場合には、S30に進む。一方、NOの場合には、S29でj=j+1し、S26に戻り、次のクラスタCmを作成することを繰り返す。

【0045】S30は、i=(全属性数-1)か判別する。YESの場合には、S32に進む。一方、NOの場合には、S31でi=i+1し、S25に戻り繰り返す。S32は、Pmax=0、かつCmax=0と初期設定する。

【0046】S33は、i=0と初期設定する。S34は、j=i+1する。S35は、クラスタCiとクラスタCjに共通して含まれる事例数|Ci ∩ Cj|を数える。

【0047】S36は、CiとCjが独立な場合の推定共通事例数eijとその標準偏差dijを計算する。図7のS37は、|Ci ∩ Cj| > eij + K · dijか判別する。YESの場合には、S38に進む。NOの場合には、S40に進む。

【0048】 S38は、 $P_{max} < (|C_i \cap C_j|) / (|C_i| + |C_j|)$ が判別する。YESの場合には、S39に進む。NOの場合には、S40に進む。S39は、 $P_{max} = (|C_i \cap C_j|) / (|C_i| + |C_j|)$
 $C_{max} \leftarrow (C_i, C_j)$ を行なう。

【0049】 S40は、 $j = m - 1$ が判別する。YESの場合には、S42に進む。NOの場合には、S41で $j = j + 1$ し、S35に戻り繰り返す。S42は、 $i = m - 2$ が判別する。YESの場合には、S44に進む。NOの場合には、S43で $i = i + 1$ し、S34に戻り繰り返す。

【0050】 S44は、 $C_{max} \neq 0$ が判別する。YESの場合には、S45に進む。NOの場合には、終了する。S45は、 C_{max} から C_m をオペレータにより作る。

【0051】 S46は、 $m = m + 1$ する。S47は、 $m = l$ か判別する。YESの場合には、終了する。NOの場合には、S32に戻り繰り返す。

【0052】

【発明の効果】以上説明したように、本発明によれば、各属性の各値についてその値を持つ事例を集めてクラスタを生成し、クラスタ集合中で共通して出現する事例の数の多いクラスタは関連が高いと見做して複数クラスタから新たなクラスタを生成することを繰り返してクラスタ階層を生成する構成を採用しているため、クラスタ生成時間を短くして1つの事例が複数のクラスに属することを許すクラスタ階層を生成することができる。これらにより、

(1) 属性の値毎にクラスタを1つづつ生成し、共通して出現する事例の数が多いクラスタ同士から新たなクラスタを生成し、1つの事例が複数のクラスに属するよ

うな場合にも対応することが可能となる。

【0053】 (2) クラスタ階層の生成に必要な処理時間は、階層中のクラスタ数の3乗と全事例数に比例する。通常、複雑すぎるクラス集合は、知識として利用し難いため、必要とされない。そこで、パラメタとして設定される階層中のクラスタ数を、事例数に比べて小さい値にできる。その結果、比較的短い時間でクラスタの生成が可能となる。

【0054】 (3) 上記(1)および(2)により、1つの事例が複数の概念に属するような大規模な事例集合に対して、効果的にクラスを構成するクラスタを学習することが可能となる。

【図面の簡単な説明】

【図1】 本発明のシステム構成図である。

【図2】 本発明の動作説明フローチャートである。

【図3】 本発明の説明図(その1)である。

【図4】 本発明の説明図(その2)である。

【図5】 本発明のクラスタ階層図である。

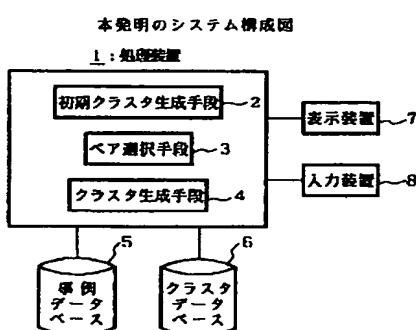
【図6】 本発明のシステム動作説明フローチャート(その1)である。

【図7】 本発明のシステム動作説明フローチャート(その2)である。

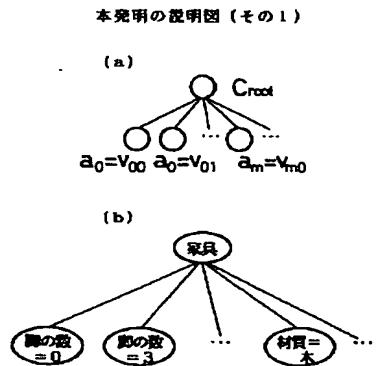
【符号の説明】

- 1 : 処理装置
- 2 : 初期クラスタ生成手段
- 3 : ペア選択手段
- 4 : クラスタ生成手段
- 5 : 事例データベース
- 6 : クラスタデータベース
- 7 : 表示装置
- 8 : 入力装置

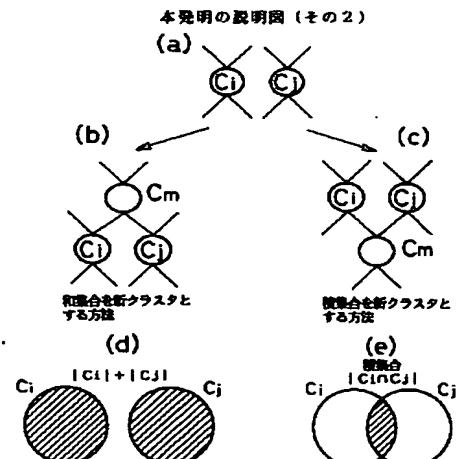
【図1】



【図3】

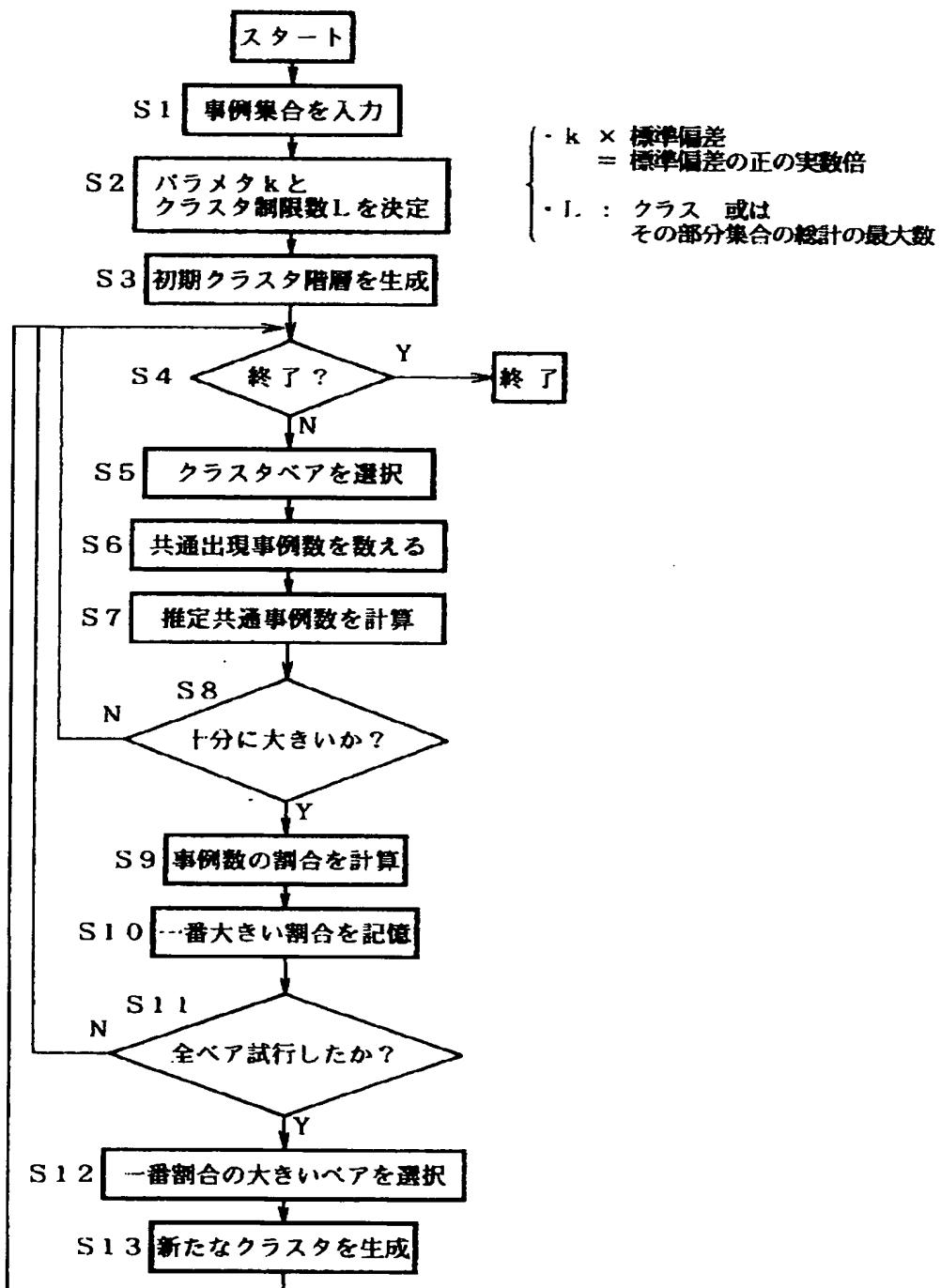


【図4】



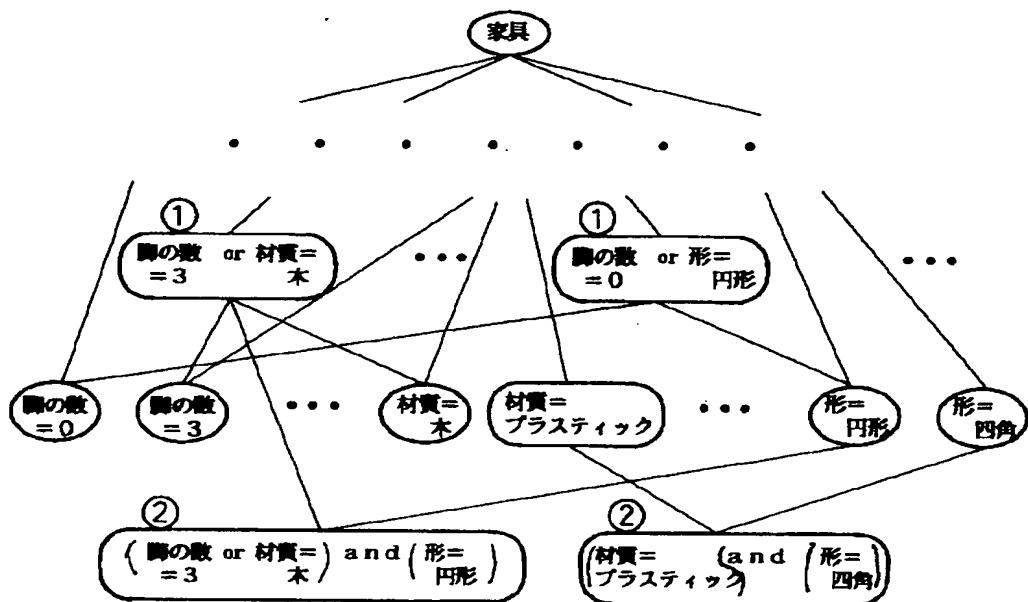
【図2】

本発明の動作説明フローチャート



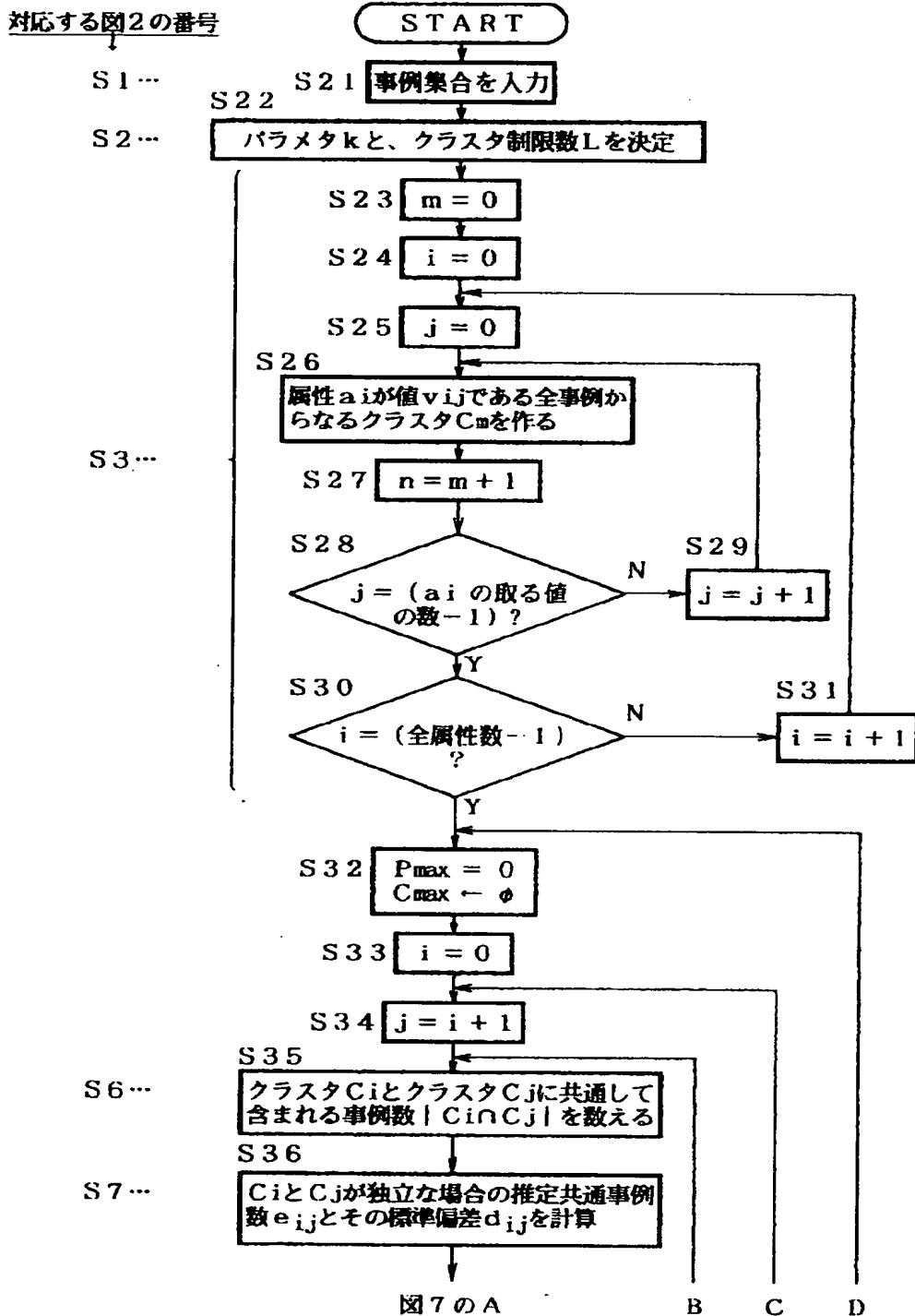
【図5】

本発明のクラスタ階層例



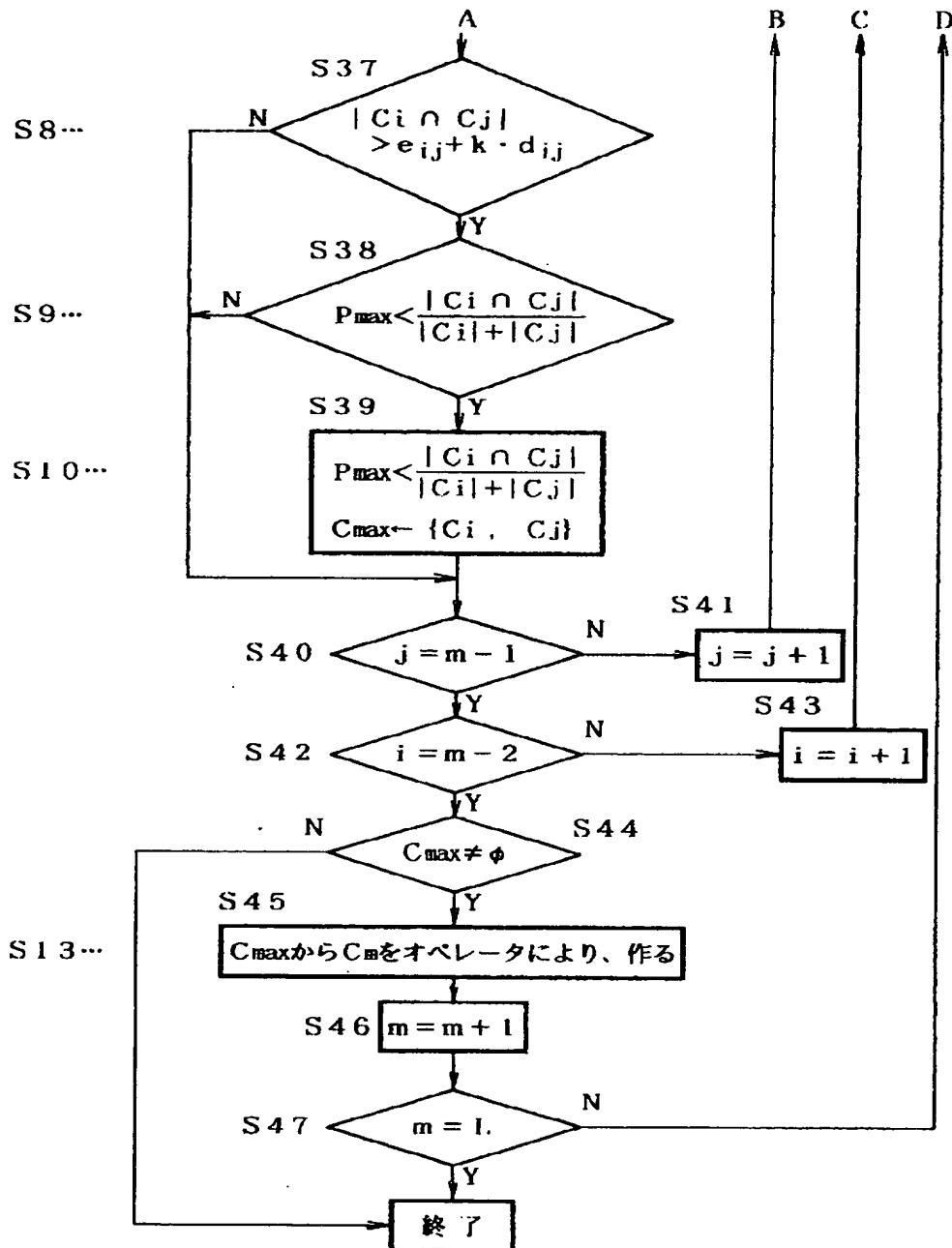
【図6】

本発明のシステム動作説明フローチャート（その1）



【図7】

本発明のシステム動作説明フローチャート（その2）



フロントページの続き

(72)発明者 岡本 青史

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(72)発明者 佐藤 理

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内